IN THE CLAIMS

1. (Currently Amended) A system for distributing connections from clients on an external network to a plurality of servers on an internal network, the system comprising:

a client interface to the external network, the client interface being operative to receive and send packets to and from a remote client;

a server interface to the internal network, the server interface being operative to receive and send packets to and from a plurality of servers, the plurality of servers being operative to establish a connection with the remote client and the system being configured to monitor connections established between the plurality of servers and clients on the external network;

a plurality of predicted responsiveness indicators, each of the plurality of predictive responsiveness indicators being associated with at least one of the plurality of servers, the predicted responsiveness indicators being operative to predict the response time of each of the plurality of servers based at least in part on response time data and aging of the response time data gathered at the system in the course of monitoring connections established between the plurality of servers and clients on the external network, the predicted response time for each of the plurality of servers being a function of the number of client connections to a particular server, the predicted responsiveness indicators being stored within the system in a manner that the predicted responsiveness indicators may be accessed; and

a predicted responsiveness comparator which is operative to access and compare the predicted responsiveness indicators and to determine which servers from among the plurality of servers is associated with a predicted responsiveness

indicator which measures a best predicted response time, the predicted responsiveness comparator being further operative to select a pointer to a server which has a predicted responsiveness that is the best predicted responsiveness among the predicted responsiveness of the plurality of servers;

whereby the server which has a predicted responsiveness which is the best predicted responsiveness is selected to handle the next connection from a client.

2. (Original) A system as recited in claim 1, wherein the predicted responsiveness indicators are periodically updated.

3. (Original) A system as recited in claim 1, wherein the predicted responsiveness indicators include the number of connections to each of the plurality of servers.

4. (Canceled).

5. (Original) A system as recited in claim 1, wherein the predicted responsiveness indicators include the predicted response time of each of the plurality of servers.

6. (Canceled).

7. (Canceled).

8. (Canceled).

9. (Canceled).

10. (Canceled).

11. (Canceled).

12. (Canceled).

13. (Canceled).

14. (Canceled).

15. (Canceled).

16. (Canceled).

17. (Canceled).

18. (Canceled).

19. (Canceled).

20. (Canceled).

21. (Canceled).

22. (Currently Amended) A system for distributing connections from clients on an external network to a plurality of servers on an internal network, the system comprising:

means for receiving and sending packets to and from a remote client;

means for receiving and sending packets to and from a plurality of servers, the plurality of servers being operative to establish a connection with the remote client;

means for monitoring connections established between the plurality of servers and clients on the external network, the means for monitoring connections comprising means for gathering response time data at the system in the course of monitoring connections between the plurality of servers and clients on the external network;

means for predicting the response time of each of the plurality of servers based at least in part on response time data and aging of response time data gathered at the system in the course of monitoring connections established between the plurality of servers and clients on the external network, the predicted response time for each of the plurality of servers being a function of the number of client connections to a particular server; and

means for comparing the predicted response time of each of the plurality of servers to select a pointer to a server which has a best predicted response time of the plurality of servers;

whereby the server that has the best predicted response time is selected to handle the next connection from a client.